Jueves 20 y Viernes 21 de Noviembre de 2014
International Conference on Regional Science:
FINANCING AND THE ROLE OF THE REGIONS AND TOWNS IN ECONOMIC RECOVERY
Facultad de Economía y Empresa y Paraninfo de la Universidad de Zaragoza
Universidad Zaragoza
Facultad de Economía y Empresa Universidad Zaragoza
ersa
RSAI

Urban agglomerations effect over the household cost of living. An analysis for the Spanish case.

Elena Lasarte Navamuel

Esteban Fernández Vázquez

Fernando Rubiera Morollón

REGIOlab - University of Oviedo, Oviedo (Spain)

Abstract

The effects of the urban agglomerations over the productivity, income, wages and many other socio-economic variables are widely studied in the literature. There are also many analyses of the effects of large cities over the prices. In line with this pervious research the objective of this paper is to measure, for the case of Spain, until what degree the cost of living could be affected by urban agglomerations. Increments in prices observed in largest cities do not necessarily imply lower household costs of living because families could adapt their purchase basket using the greater diversity of products maintaining their utility level. According with this idea we propose use micro-data of the Household Budget Survey of the Spanish Institute of Statistics to calculate a household true cost of living consistent with the microeconomic foundations. A fixed utility approach is used instead of a fixed basket one for each family. A Quantile Regression procedure is used in other to identify different factors which influence the cost of living of the household, especially the agglomeration factor, across the distribution. The results when the characteristics of the families are controlled show that differences in household cost of living of large cities is even greater than the one observed with the simple household costs of living aggregations by city-size. This is especially clear in the upper quantiles of the distribution.

Key words: *Cost of Living (COL), Almost Ideal Demand Systems (AIDS), Quantile regressions, household consumption, city-size and Spain.*

JEL Classification: *D12, R11 and R22.*

1. Introduction

The concept of *agglomeration economies*, first proposed by Weber (1909), is central in Regional and Urban Economics. Ohlin (1933), Hoover (1937) and Isard (1956) clarify the idea and distinguish different types of *agglomeration economies*: (i) *economies of scale*, (ii) *localization economies* and (iii) *urbanization economies*. *Economies of scale,* internal to a firm, are related with the concentration of population in an area that means that bigger market sizes provoke the possibility of having lower production costs. *Localization economies*, also known as Marshallian economies, are external to the firm but internal to an industry and are the set of positive externalities produced by the concentration of similar firms in a reduced area. Finally, *urbanization economies* are the externalities, sometimes positive and others negative (*diseconomies*), derived from the spatial concentration of both firms and population which are a fundamental ingredient for understanding the link between city size and income per capita or productivity, the distribution of economic activity across space, the importance of cities in the economic growth of regions and countries; not to mention international and interregional trade, industrial location, cluster formation or regional specialization.

If *agglomeration economies*, particularly the *urbanization economies*, are so relevant in the explanation of so many economic behaviors it would seem logical that they are also a fundamental concept to understand the spatial dynamics of others aspects such as consumption patterns and price dynamics. There are previous empirical studies that suggest that the place of residence, in an urban or rural environments or even the size of the city, affect the consumption behaviors. Large cities offer a greater variety and higher quality of goods attracting people with particular characteristics and generating different styles of life. As a result, the consumption patterns generated in the metropolis are different than those generated in small cities or rural areas. Moreover, is also hopped that land pressure and local amenities found in metropolitan areas make prices to be higher. But from this is not possible to deduce that the higher prices or different style of life of cities will increase the cost of living. Families could use the greater variety of products and options that the large city offers to maintain their standard of life (utility level) with higher prices or under different conditions.

Our objective in this paper is to explore empirically this issue measuring if *urban agglomerations* could increase families' cost of living in the specific case of Spain. This country is especially interesting because the urban system is very complete containing big metropolitan areas, several medium-size cities with different economic structures and geographical characteristics, all surrounded by an important extension of rural areas.

There are many empirical works that support disparities in prices, consumption patters and cost of living among metropolitan areas and regions, the majority of them are for the US (Haworth and Rasmussen, 1973; Cebula, 1980 and 1989; Hogan, 1984; Walden, 1998; Kurre, 2003; and Cebula and Todd, 2004). Another little work has been done in Europe; Hayes (2005) estimates UK regional price indices for 1974 to 1996 finding more regional price variations than variations over the whole sample period. Kosfeld et al. (2008) and Blien et al. (2009) evidence regional cost of living differences in Germany for different purposes. But in Spain there is no evidence of such studies. In any case, all these quoted researches use as a measure of prices and cost of living some kind of official indexes, like Consumer Price Index (CPI), which normally evaluate changes in the average prices for the acquisition of a fixed basket of goods considered as representative of all consumers, ignoring the fundamental consumers' substitution because of changes in their preferences or adaptation of their consumption decisions to the residential characteristics. Consequently these indices do not reflect the "true" cost of living.

The theory of the "true" cost of living (Könus, 1939) establishes that a "true" cost of living must be consistent with the microeconomic foundations and must recover the differences in preferences among consumers. This is possible using a fixed utility approach instead of a fixed basket one, this means that fixing the utility level, a "true" cost of living measures the cost of attaining a utility level at given prices. Since utility levels data are not available from National Statistical Agencies it is necessary to estimate it using the Almost Ideal Demand System (AIDS) procedure of Deaton and Muellbauer (1980). This approach will allow us to calculate an indicator of cost of living at the same utility level something that is especially relevant in the quantification of differentials in the costs of living between large and small cities, since people living in larger cities benefit from a greater variety of goods that enhances the substitution in their consumption.

To apply this procedure to the Spanish case we are going to use the Household Budget Survey (HBS) of the National Statistical Institute (INE). Additionally, we are going to work with the maximum level of disaggregation: the household level. Instead of using any official aggregated price index we are going to estimate a "true" cost of living for each family. The advantage of working at a micro level is that the more disaggregated cost of living allows us to isolate the model from the factors inherent to the households and to the individuals focus the attention in the pure effect of the agglomerations. But although it seems the ideal framework, this approach is very complicated and it most of the cases is not operational due to the data requirements. For this reason, we calculate the micro cost of living using *unit values* only for the food group, due to this group, together with the group of energy, the only one that reports the necessary data to calculate the cost of living. This limitation is not a big shortcoming because as Slesnick (2002) pointed to, differences in price levels are obvious in goods such as housing, but the critical question is whether the dispersion in other representative consumer goods is pervasive and of sufficient magnitude to influence the costs of living of households significantly.

Once we have a "true" cost of living at household level we can aggregate this information according with the city-size and we can observe if the costs supported by families located in large urban areas are significantly higher than those that are supported by families in small cities or rural areas. Nevertheless, although this could give us a first intuition of how urban agglomerations affect the standard of living we should consider that the possible differences could be explained by processes of concentration of families with higher income or with different consumption behaviors. To delimitate the exact effect of urban agglomerations on the cost of living we must control for household and regional characteristics identifying the specific effect of the urban agglomerations. The second contribution of the paper is that this simple model of determinants of the household cost of living will allow us to identify the specific role of urban agglomerations. In the empirical estimation of this model a Quantile Regression procedure is used. This method not only allow us to know how the determinants include in the model influence the cost of living, but for whom these determinants influence more.

The paper proceeds in the following way: in section 2 a brief review of the theory of the “true” cost of living indices will be provided. In section 3 it will be explained the methodology used for applying the Konus (1939) theory for the Spanish case and show the first results. Section 4 recovers the model of the determinants of the cost of living and described the results obtained. And, finally, section 5 summarizes the main findings of the paper.

2. The “true” cost of living: a brief review

The theory of the “true” Cost of Living (COL) was first developed by Konüs (1939) who defined the COL as the monetary value of the goods consumed in a period by a household which are necessary for the maintenance of a certain standard of living. The “true” cost of living was originally proposed to measuring the differences on the cost of living along the time. It has also successfully extended to study the differences across space, Spatial Cost of Living (SCOL) using the same basic idea but comparing two points in space (Desai, 1969; Nelson, 1991; Timmins, 2006 and Atuesta and Paredes, 2012). Thus in computing a “true” COL it is compared the monetary cost of two different combinations of goods which are connected solely by the condition that, during the consumption of the two combinations, the standard of living provided by both is exactly the same.

However, the usual method of calculating the “true” COL is the so-called method of aggregates. It consists on calculate the cost of a given basket of goods corresponding to the average or normal consumption and at prices prevailing at a given time, and dividing it by the cost of the same basket of goods at prices of another period. But this method does not show exactly the “true” COL because there is the assumption that while prices change consumption does not change. But, in reality, consumers change its consumption due to rises and falls in prices in order to maintain its standard of living.

In order to construct a “true” COL it is necessary to know which combination of goods yields a given standard of living despite price changes. For this purpose it is used the concept of *indirect utility function,* the consumer is going to maximize its utility function at a given prices and subject to a budget restriction. The formulation of the COL would be:

$$\boldsymbol{COL = c(p, u)} \qquad [1]$$

where p are prices faces by consumers, where u is he utility function to be reached by the consumer, and c is the cost of attaining the utility level u at prices p.

The major problem arises from the unknown and not observable utility function, and without knowing the utility function is impossible to derive the cost function and to calculate the COL. The typical solution to address this problem is to follow a flexible function demand system with several convenient properties. These flexible functional forms permit the estimation of demand equations without knowing explicitly the functional form of the utility function. The flexible functional form to be used in this research will be the Almost Ideal Demand System (AIDS) proposed by Deaton and Muellbauer (1980).

The point of departure for estimating an AIDS starts by defining a PIGLOG class cost or expenditure function, a special case of the Price-Independent Generalized Linear cost function, proposed by Muellbauer (1975) and consistent with the microeconomic theory that sets the minimum expenditure necessary to attain a specific utility level at given prices for a set of n products:

$$\log c(p, u) = (1 - u) \log\big(a(p)\big) + u \log\big(b(p)\big) \qquad [2]$$

where c is the expenditure function, p is the price vector and u is the utility level. With some exceptions, u lies between 0 (subsistence level) and 1 (bliss level) so $\log\big(a(p)\big)$ and $\log\big(b(p)\big)$ can be considered as the log of the costs of subsistence and bliss, respectively. Their respective functional forms are:

$$\log\big(a(p)\big) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \log p_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log p_i \log p_j \qquad [3]$$

$$\log\big(b(p)\big) = \log\big(a(p)\big) + \beta_0 \prod_{i} {p_i}^{\beta_i} \qquad [4]$$

where the i sub-index (j) denotes the products included in the demand system. The demand functions can be derived substituting [3] and [4] in the cost function [2], from which we obtain:

$$\log c(p,u) = \alpha_0 + \textstyle\sum_{i=1}^{n} \alpha_i \log p_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log p_i \log p_j + u\beta_0 \prod_i {p_i}^{\beta_i} \qquad [5]$$

By applying the Shepard's lemma to [5], i.e., price derivatives are equal to the quantities demanded, and multiplying both sides of the equation [5] by $p_i/c(u,p)$, we obtain:

$$\frac{\partial \log\big(c(p,u)\big)}{\partial \log p_i} = \frac{p_i q_i}{c(p,u)} = w_i \qquad [6]$$

where w_i is the budget share of good i:

$$w_i = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \log p_j + \beta_i u \beta_0 \prod_i p_i^{\beta_i} \qquad [7]$$

To obtain an estimable system we need to solve for u as a function of observed and known parameters from equation [5]:

$$u = \frac{\log c(u,p) - \alpha_0 - \sum_{i=1}^{n} \alpha_i \log p_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log p_i \log p_j}{\beta_0 \prod_i {p_i}^{\beta_i}} \qquad [8]$$

Substituting u in equation [7] we obtain:

$$\begin{aligned} w_i = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \log p_j + \beta_i(\log c(p,u) - \alpha_0 - \sum_{i=1}^{n} \alpha_i \log p_i \\ - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log p_i \log p_j) \end{aligned} \qquad [9]$$

The shares in [9] are determined by prices and the expenditure function, plus a set of parameters to be estimated. These shares are the AIDS demand functions and they can be expressed as:

$$w_i = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \log p_j + \beta_i \log\{x/P\} \qquad [10]$$

where α, β and γ are the parameters to be estimated, x is the total expenditure on the and P is a price index defined as:

$$log\,P = \alpha_0 + \sum_{j=1}^{n} \alpha_j\, log\, p_j + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij}\, log\, p_i\, log\, p_j \qquad [11]$$

Some empirical studies use the Stone Price Index to avoid problems of non-linear estimations. However, we estimated the original model as suggested by Deaton and Muellbauer (1980) using the TRANSLOG price index described in [11][1].

The parameters included in the AIDS model should satisfy a set of constrains. Firstly, they must hold the adding-up restriction ($\sum_{i=1}^{n} w_i = 1$), which requires equality of the sum of individual commodity expenditures and the total expenditures:

$$\sum_{i=1}^{n} \alpha_i = 1, \quad \sum_{i=1}^{n} \gamma_{ij} = 0, \quad \sum_{i=1}^{n} \beta_i = 0 \qquad [12]$$

Furthermore, the equations of the AIDS are homogeneous of degree zero in prices and total expenditure taken together. This means that if prices and total expenditure increase by the same amount the demand remains unchanged:

$$\sum_{j=1}^{n} \gamma_{ji} = 0 \qquad [13]$$

Moreover, the total expenditure must verify the Slutsky symmetry, which requires that the compensated cross-price derivative of commodity *i* with respect to commodity *j* equals the compensated cross-price derivative of commodity *j* with respect to commodity *i*:

$$\gamma_{ij} = \gamma_{ji} \qquad [14]$$

The β and γ parameters can be interpreted in economic terms. The γ_{ij} elements quantify the effect of changes in relative prices, representing the % of change on the i_{th} budget share produced by a 1% increase in the price of the j_{th} product, being (x/P) held constant. The effects of changes in the real expenditure operate

[1] As an alternative to [12], Cooper and McLaren (1992) suggest a modification of AIDS called MAIDS, which preserves regularity in a wider region of the expenditure-price space. Nevertheless, the most usual form in the literature is AIDS or its linear approximation, LAIDS.

through the β_i coefficients, which are positive for luxuries and negative for necessities (Deaton and Muellbauer, 1980).

3. Estimation of a "true" cost of living at household level: application and first results to the Spanish case

Both in the time dimension and the spatial context the researcher is comparing aggregated information at regional or national units. The proposal for this paper is working at the maximum level of disaggregation: the household level. The idea is to estimate a household "true" cost of living (COL) for Spain in order to analyze the determinants of the costs of living focusing in the role of agglomerations over these costs of living.

As in many other countries, the application of this approach to the Spanish case entails the difficulties arising from the lack of available data. The only survey that contains information on household expenditure and consumption patterns is the Household Budget Survey (HBS), an extensive survey of Spanish household purchases, income and other socioeconomic characteristics with 21,790 observations. The Spanish Statistical Institute (INE) conducts this survey annually with different households every year. The estimation of the AIDS requires information on prices, quantities purchased and expenditures at the household level. As all the prices must be observable to estimate the model, the unitary values at which households purchase the commodities are recovered by dividing expenditures by quantities[2]. All these information requirements limit the estimation to be feasible only for the food group, being the only type of product studied in the HBS with detailed information about the variables required. The data of these products are classified into ten food sub-groups: (i) *Bread and cereals*, (ii) *Meat*, (iii) *Fish*, (iv) *Milk, cheese and eggs*, (v) *Oil*, (vi) *Fruits*, (vii) *Vegetables*, (viii) *Sugar*, (ix) *Coffee, tea and cacao*; and (x) *Mineral water and soft drinks*.

For each group $i = 1, \dots, 10$ the observed budget share w_i of equation [10] in each household is calculated by dividing the expenditure of the household in this specific sub-group by the total household expenditure in food.

[2] This procedure to obtain the unit prices is accepted in the literature and it is well known as *unit values* (Deaton, 1988).

An additional issue in the estimation process, derived from the characteristics of the HBS, is that prices are not available for all items in all households. This situation happens when a household does not really consume that specific group, being the consequence that the price of the item cannot be recovered by means of unit values. This issue provokes that the dependent variable is *truncated* or *censored*. For solving this problem the price of the item has been replaced by a geometric mean of the prices of this item in the same region[3], distinguishing the kind of municipality where this item was been purchased. In these cases, the price is replaced by the average price of the same item in the same region and in the same kind of city.

The model to be estimated in our case is a specific version of the AIDS model where censored data and spatial factor are considered. The modeling of demand systems with household-level microdata has the advantage of providing a large and statistically rich sample avoiding the problem of aggregation over consumers. In the other hand, detailed microdata may cause a problem of censored commodity purchases, especially when a very detailed classification for the commodities is used. Not accounting for the zero consumption biases the estimation of the parameters of the model and it may produce a selection bias if we do not incorporate these observations into the estimation process. Dealing with censored data is more complicated in the case of demand systems than in a case of the econometric estimation of one single equation. The complication arises from the necessity of ensuring nonnegative estimates of the quantities consumed; the requirement of including the constraints imposed by economic theory; and the numerical problem of having to evaluate high-dimension cumulative density functions during the estimation (Dong *et al.*, 2004).

To address these problems we will follow the two-step method proposed by Shonkwiler and Yen (1999), which improves the previous "favorite" two-step estimation procedure of Heien and Wessells (1990). In the first step we estimate a PROBIT regression with a dependent binary variable that represents the household decision of consuming or not, which takes the value of 1 if the household purchases the commodity and the value of 0 if not, which depends on a

[3] This is a usual procedure to replace prices that are missing, Dong *et al*. (2004) and Atuesta and Paredes (2012) use the same procedure for Mexico and Colombia, respectively.

set of socioeconomic variables that are used as regressors. The PROBIT model determines the probability that a given household consumes a given good and it is used to estimate the cumulative distribution function (Φ) and the normal density function (ϕ). The second step includes the cumulative function $\Phi(x)$ as a scalar in the equations for shares, while the density function $\phi(x)$ is included as an extra explanatory variable.

In this case the AIDS model to be estimated is of the form:

$$w_i = \Phi(x)\left[\alpha_i + \sum_{j=1}^{n} \gamma_{ij}\, log\, p_j + \beta_i\, log\{x/P\}\right] + \sum_k c_k\, CS_k + r_h R_h + \delta_i \phi(x) \qquad [15]$$

where δ_i is a parameter associated with the density function, CS_k are dummy variables for different urban sizes and R_h is a regional dummy for each one of the NUTS-II regions of Spain, and c_k and r_h are the parameters associated with each type of dummy, respectively, with the aim of recover the idiosyncratic components inherent to each region and type of city. The estimation of the parameters is made by applying Nonlinear Seemingly Unrelated Regression (NLSUR), which estimates a system of nonlinear equations by Feasible Generalized Nonlinear Least Squares (FGNLS). With the parameters estimated we recover the expenditure functions for each household defined as in Equation [16]:

$$log\, c(p,u) = \alpha_0 + \textstyle\sum_{i=1}^{n} \alpha_i\, log\, p_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij}\, log\, p_i\, log\, p_j + u\beta_0 \prod_i {p_i}^{\beta_i} \qquad [16]$$

The $log\, c(p,u)$ represents the COL for each household in Euros needed to attain the median utility level of the country as a whole. More precisely, this COL is calculated with the prices faces by each household, with the expenditure level of each household applying the median utility level of the country.

Before to present our model of the determinants of the COL, are showed the "true" cost of living calculated at a household level for Spain in 2012. In the Table 1 is summarized the COL by percentiles and distinguishing if the household resides in an agglomeration that is a city of more than 100,000 inhabitants, or not.

Table 1 Cost of living in Euros by percentiles in 2012 in agglomerations *vs.* non agglomerations

	Mean	10	25	50	75	90
>100,000 inhabitants	3692.82	2589.14	3043.09	3593.42	4217.34	4848.84
<100,000 inhabitants	3501.68	2434.87	2859.88	3409.17	3997.47	4633.77
% Difference	5.46%	6.33%	6.41%	5.40%	5.50%	4.64%

Results in this Table 1suggest that the smallest areas benefit from reduced costs of living when compared with the largest cities of Spain. The estimates of cost of living by city size seem to be coherent with the expectations about the effects of agglomeration economies in recent literature indicating that the largest cities have suffered the highest cost of living all along the period under study, being the smallest cities the areas where these estimates get the lowest values. These differences on average range from around more than 5.46% in 2012, suggesting that the higher market competition and the wider variety of products present in large cities are not enough to offset the spatial competition and land pressure that characterize these big cities.

4. Analysis of the determinants of the differences in the cost of living among the Spanish families

Previous section results show higher costs of living in Spanish large urban agglomerations, which implies lower standard of life in those places. But, the question that all the previous literature cannot solve is which part of this increment in the cost of living is due to a process of agglomeration of qualified professionals and richer families in main urban areas and what is just due to the city size effect over the type of consumption, prices and, consequently, cost of living. Because the biggest cities attract a particular population with particular characteristics this make that a particular consumption patterns take place in agglomerations, in turn, this consumer behavior makes that the demand of certain goods rises exerting a pressure over prices of all goods and services. This process makes that agglomerations are more expensive to live in. But at the same time it could be observed that individuals with the same characteristics have a different consumption behavior by the fact those agglomerations promote a particular consumption which is not found in small areas.

The aim of this section is to develop a model capable to explain the determinants of the cost of living in a place. The micro-cost-of-living will be regressed, through a Quantile Regression, over several variables to isolate the pure effect of agglomerations over the cost of living by controlling for individuals' and households' characteristics. In order to examine the determinants of the cost of living (COL) variation among the Spanish families we postulated a model of COL determinants focus our attention in the effect of agglomerations over this COL.

4.1. An empirical model to estimate the household cost of living determinants

Basic economic theory could be used to find the determining factors of the cost of living variations. As Kurre (2003) explain, the fundamental idea is that factors that increase the demand of goods cause prices to be higher; those which tend to increase supply cause prices to be lower. Additionally, there exist idiosyncratic factors of a region which can influence the cost of living, for example the climate conditions or the geographic situation in the country.

Based on this, the key variables examined are: a dummy variable which represents if the household belongs to a city of more than 100,000 inhabitants; income per capita in the Autonomous Community; one dummy for each region of the country at NUTS-I level; and a set of variables representing various characteristics of the household, like the size of the household, the number of employed, the number of dependents; and of the household head, like the age, the income level and the level of education. The latter variables which compose the vector Z in equation [18] are include as control variables to try to isolate the pure effect of the size of the city over the COL. These variables are expected to have the effects that predict the consumer theory.

Previous analyses in the empirical literature have also demonstrated the strong relation between income and costs of living. The low income areas have the lowest COL and the high income areas have the highest one, in general, the richer the area, the higher the demand for goods, so the higher the pressure on prices. This relationship is found strongly remarkably in works such as Hogan and Rex (1984), McMahon (1991), Kurre (2003) and Kosfeld *et al*. (2008).

The influence of geographic variations over the cost of living is also well documented. In Hogan (1984) is revised some empirical works in this issue, for example, Shefer (1970) and Sherwood (1975) evidence highest cost of living in the North East and lower in the South; and Haworth and Rasmussen (1973) found lower living cost in the South. Gradually, more evidences have emerged; McMahon and Melton (1978) and McMahon (1991) concluded that the Southern US benefits from lower costs of living compared to the Eastern Seaboard and the Northeast. In Europe, Hayes (2005) found a great impact of regional price variations in the South East Region of the UK; Kosfeld *et al*. (2008) find strong evidence for the presence of spatial price effects using Consumer Price Index for the Bavarian districts. In this work we also hope to find remarkably differences between the regions included in the model, this regions are included in form of a dummy variable, one for each region (NUTS-I) that is Northwest, Northeast, Region of Madrid, Central Region, East Region, South Region and Canary Islands.

Is not immediately clear the effect of the agglomerations over the cost of living, the magnitude of the city's population could affect the cost of living in at least three magnitudes (Haworth and Rasmussen, 1973): (i) economies or diseconomies of scale in the provision of public services; (ii) externalities affecting the compensation of those employed in the city; and (iii) the cost of land. On the one hand, if there is more population the demand of the goods rise and, consequently, the price of the goods rise too. But, on the other hand, large population can produce economies of scale in the production process and lead to lower prices. Cebula (1980 and 1989) finds that the second factor predominates over the first one, so the more the population, the lower the cost of living. In contrast, other authors like Blien *et al*. (2009) find that larger cities are more expensive to live in. In the model proposed here is introduced a dummy variable which represents the agglomeration effect, this dummy variable is that of cities of more than 100,000 inhabitants. The reason for choosing this is because in the previous section it could be seen that the most striking differences took place between these cities and the rest ones.

The variables can be simplified as:

$$COL_i = f(Agglomeration, X, Z) \qquad [17]$$

Where *Agglomeration* is a dummy variable that represents the cities of more than 100,000 inhabitants, X is a set of geographic and regional variables relating to each region at which the households belong to; and, *Z* is a set of households' and individuals' characteristics variables. In the Table 2 are summarized the variables of the model. The main variable in our model, which is *Agglomeration* has been defined as we described in the lines below due to the data restrictions. The HBS used in this research only permit us identify five types of cities: cities of more than 100,000 inhabitants; cities between 50,000 and 100,000; cities between 20,000 and 50,000; cities between 10,000 and 20,000; and rural areas of less than 10,000 inhabitants. For this reason, it has been consider as agglomerations the cities of more than 100,000 inhabitants and it is going to be studied the effect of those in contrast to the rest of types of cities smaller than the agglomerations considered.

Table 2 Description of the variables of the model of eterminants of *COL*

Variables		**Source**
Agglomeration	Dummy variable that represents the cities of more than 100,000 inhabitants	HBS
Vector X of regional characteristics	Income per capita in the Autonomous Community	Regional Accounts
	Dummy variables for each region at the NUTS-I level	HBS
Vector Z of household characteristics	Household size Number of employed Number of dependents Age of the household head Education of the household head Income level of the household	HBS

The dependent variable of the regression [17] is the Cost of Living (COL) at the individual level provided by our own estimations. Once the COL for the 21,484 households is calculated it is proceeded to estimate the full regression [18] for quantiles 1-99:

$$Q_\theta[COL|Agglomeration, X, Z] = \beta_\theta^0 + Agglomeration\beta_\theta^1 + X\beta_\theta^2 + Z\beta_\theta^3 \qquad [18]$$

where COL is the log of the Cost of Living in Euros of each household, $Q_\theta[COL|X,Z]$ is the θth conditional quantile of COL, β_θ^0 is the regression intercept, $Agglomeration$ represents the cities of more than 100,000 inhabitants, X and Z are covariates matrix which include all geographic and household regressors, respectively, and, the coefficients β_θ represent the returns to covariates at the θth quantile.

The process yields a sample of 21,484 observations. The intercept X recovers the Income per capita of the Autonomous Community of the household and the set of dummies of Spanish regions; the intercept Z recovers the Household Size measured as the number of members of the household, the Number of Employees in the household measured in number of people employed, the Age of the household head measured as a continuous variable that represents the number of years old, the Number of Dependents in the household, Education Level of the household head which is divided into four categories: no studies, first cycle studies, second cycle studies and high degree studies, and the income level of the household which is divided into seven categories which range from less than 500 net Euros per month to more than 3,000 net Euros per month.

With the described model it is estimated a quantile regression model (Koenker and Basset, 1978) which fits quantiles to a linear function of covariates. In its simplest form, the least absolute deviation estimator fits medians to a linear function of covariates. The method of quantile regression is more attractive because medians and quartiles are less sensitive to outliers than means, and therefore Ordinary Least Squares (OLS). Indeed, the likelihood estimator is more efficient than the OLS one. Quantile regressions permit that different solutions at different quantiles may be interpreted as differences in the response of the dependent variable to changes in the regressors, thus, quantile regressions detect asymmetries in the data which cannot be detected by OLS. But the most important feature is that quantile regression analyzes the similarity or dissimilarity of regression coefficients at different points of the dependent variable, which in this case is the household "true" COL; it allows one to take into account the possible heterogeneity across COL levels. The model is estimated in using the least-absolute value minimization technique and bootstrap estimates of the asymptotic variances of the quantile coefficients are calculated with 20 repetitions.

4.2. Main results

Table 3 gives us the results of the OLS (first column) and Quantile Regression estimations (rest of the columns) of the households cost of living calculated in previous section as a function of the regional and the socioeconomic variables described above. The first column of Table 3 gives the results of the OLS regression, the successive columns gives the results of the 10, 25, 50, 75 and 90 quantiles, respectively.

We can observe that in both cases, with OLS or Quantile Regression procedure, almost all variables are significant at 1% level except a few. If we regress the same variables with the expenditures of the families provided by the HBS, instead of the cost of living that were have calculated, the results are completely different: see Appendix I in which the same analysis is made but using the household expenditure as the dependent variable in which only a few variables are significant. This difference in the results with “true” cost of living and expenditure level confirms the idea that the expenditure of the families is not a proper way for measuring the effects of different factors, including the size of the city, over the standards of living due to the fact that the families could adapt their consumption to the characteristics of the place in which they are living, maintaining or increasing their utility but without changes in the expenditure.

Returning to the results of Table 3 and if we focus our attention in the first column, OLS procedure, we can observe how income per capita for each region (Autonomous Community) and the regional NUTS-I dummies are both statistically significant. The income variable represents the income per capita of the Autonomous Community at which the household belongs to. This variable is one of the most statistically significant showing a positive relationship between the income per capita and the cost of living of the household. Thus, the strong theoretical response of prices in income is supported by the data. Regional dummy variables are represented at the level of NUTS-I. The omitted region is the Autonomous Community of Madrid, so the results are interpreted respect to this region. As we can see all regional dummies are statistically significant, the

Northwest and Central dummies are negative and statistically significant; this means that living in those regions is cheaper than in the Autonomous Community of Madrid. The rest of the dummies are positive and statistically significant meaning that the cost of living in these regions is higher than in Autonomous Community of Madrid.

The Northwest and Central regions include Autonomous Communities all of them with lower costs of living than Madrid, these Autonomous Communities are Galicia, Asturias and Cantabria in the Northwest; and Extremadura, Castile Leon and Castile La Mancha in the Central region. In contrast, the rest of the regions have higher cost of living than Madrid, this can be explained by the fact that the Northeast region is formed by some of the richest Autonomous Communities that is Navarra and Basque Country. In the same way the East region is influenced by Catalonia which has a COL in 2012 5.7% higher than Madrid (Lasarte et al., 2012); the South region includes Autonomous Communities very touristic like Murcia and the Mediterranean side of Andalusia which make arise the COL respect to Madrid. Lastly, the particular position of the Canary Islands makes that the cost of living is remarkably higher than in Madrid mainly due to transportation costs.

The household socioeconomic characteristics are also significant and have the expected effect over the cost of living. The household size, number of employed, the age and number of dependents are continuous variables. The level of education is represented with a set of dummy variables that indicate the effect of each degree of studies respect to individuals which have no studies or have basic studies. Regarding with the income level the results are reported respect to the households which have less than 500 Euros of net monthly income.

The variable in which we focus our attention is the agglomeration dummy which takes value 1 if the household is located in a city of more than 100,000 inhabitants and 0 otherwise.

Quantile Regression, the rest of the columns of Table 3, gives us valuable information about for whom the effects are more relevant. In general, the results are very similar for most of the variables in all the distribution. It is not observable any relevant change in household variables. Just some differences can be observed in the effect of the regional income which is a bit superior in upper percentiles. But the effect of the agglomeration variable change significantly along the quantile distribution. The variable is statistically significant and positive in the upper budget level that is in 50, 75 and 90 percentiles, this means that the COL is higher in the biggest cities only for the rich.

This result has sense because there are some kinds of goods which are only available in the biggest cities and are only consumed by high income households. Consequently, the biggest cities have a greater demand of the goods with income elastic demands which are only demanded by rich households and this cause an upward pressure on prices. In contrast, the price of inferior goods which composed the basket of the poor, are not affected as much as the price of superior goods. In other words, the poor will never consume superior goods and their basket of goods costs similarly in all city sizes. It can be seen graphically the evolution of the coefficient of the agglomeration variable in Figure 1.

Figure 1 Evolution of the Agglomeration coefficient along the quantile distribution

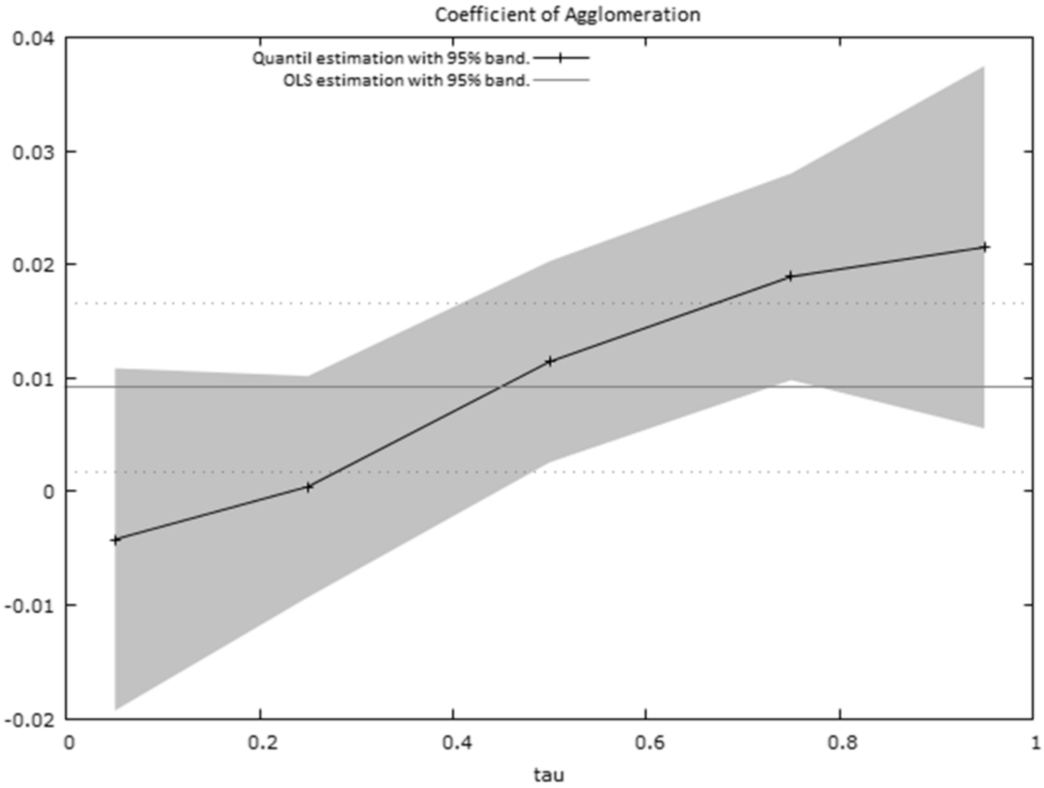

Table 3 Estimates of the OLS and Quantile Regression with the COL estimated at household level

	OLS		QUANTILE REGRESSION									
			10		25		50		75		90	
COL	Coef.	T	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Cons	5.9771***	34.77	5.6481***	20.53	6.0452***	26.8	5.9324***	30.96	5.9494***	29.82	6.1096***	34.78
Agglomeration	0.0096***	2.72	-0.0014	-0.25	0.0035	0.91	0.0103**	2.39	0.0159***	3.53	0.0211***	3.61
Income	0.2013***	12.03	0.1995***	7.35	0.1758***	7.66	0.2065***	10.69	0.2231***	11.55	0.2205***	13.03
Northwest	-0.0519**	-5.54	-0.0115	-0.65	-0.0362**	-2.45	-0.0530***	-4.38	-0.0614***	-5.49	-0.0685***	-5.05
Northeast	0.0154***	2.16	0.0465***	2.99	0.0385***	3.12	0.0188**	2.16	0.0003	0.03	-0.0116	-0.72
Central	-0.0580***	-5.53	-0.0528*	-2.85	-0.0556***	-3.47	-0.0479***	-3.56	-0.0490***	-3.3	-0.0532*	-3.3
East	0.0709***	8.8	0.0845***	5.63	0.0870***	6.45	0.0778***	7.19	0.0697***	5.77	0.0530**	4
South	0.0795***	6.96	0.0910***	4.71	0.0807***	5.4	0.0825***	6.26	0.0785***	5.29	0.0686***	4.18
Canary Islands	0.1409***	11.78	0.1846***	8.79	0.1672***	11.91	0.1444***	10.06	0.1247***	8.02	0.0893***	4.25
Household Size	-0.0162***	-7.33	-0.0184***	-7.59	-0.0245***	-13.79	-0.0242***	-10.44	-0.0187***	-6.97	-0.0109**	-2.17
Number of employed	0.0132***	4.82	0.0126**	2.66	0.0105*	2.46	0.0138***	3.35	0.0110***	4.05	0.0150**	3.7
Age	0.0004***	2.86	0.0003	1.2	0.0005*	2.49	0.0006***	3.06	0.0005*	2.24	0.0004*	2.03
Number of dependents	0.0100***	3.63	0.0173***	4.65	0.0167***	6.35	0.0155***	4.43	0.0099***	3.33	0.0019	0.31
First cycle studies	0.0138***	2.75	0.0194**	2.21	0.0232***	3.77	0.0197***	5.29	0.0122*	1.89	0.0045	0.55
Second cycle studies	0.0400***	6.57	0.0433**	4.41	0.0522***	7.93	0.0461***	6.24	0.0448***	6.04	0.0409***	4.38
High degree studies	0.0557***	9.21	0.0486***	5.23	0.0606***	11.61	0.0620***	8.46	0.0593***	7.14	0.0545***	6.87
500-1000 Euros	0.0534***	6.02	0.0765***	5.9	0.0686***	5.35	0.0441***	4.42	0.0304***	2.92	0.0400*	2.95
1000-1500 Euros	0.0788***	8.94	0.1172***	14.75	0.1118***	11.57	0.0724***	8.45	0.0495***	5.14	0.0519*	3.35
1500-2000 Euros	0.1108***	12.09	0.1526***	15.25	0.1461***	13.25	0.1042***	10.05	0.0731***	7.65	0.0724***	6.61
2000-2500 Euros	0.1339***	13.87	0.1874***	19.23	0.1793***	17.18	0.1305***	13.88	0.0907***	8.67	0.0839***	4.46
2500-3000 Euros	0.1614***	15.8	0.2190***	16.45	0.2131***	15	0.1596***	13.88	0.1158***	8.96	0.1018***	6.32
More than 3000 Euros	0.1824***	17.6	0.2504***	22.09	0.2410***	19.77	0.1848***	15.28	0.1377***	11.24	0.1228***	6.38

Note: *, ** and *** represent the level of significance to 10%, 5% and 1%, respectively.

5. Conclusions

Prices and consumption patterns change across the space. There are geographical, weather, cultural, sociological and economic reasons to offer as explanations for the fact that the level of prices and the way of consume differ from one region to another. Particularly relevant are the potential effects of the size of the cities. Large cities are more competitive, offer a greater variety of goods and services and, among other factors, develop a different style of life… As a result, the response of consumers to changes in prices should be different in a small town in contrast to a large metropolis.

Although there is ample evidence of how consumption patterns are affected by factors such as the level of income or stage in their life cycle at which households find themselves, the empirical studies on spatial effects are limited and contradictory. Several studies have found significant differences in consumption patterns of households living in rural areas compared to those residing in urban areas. However, most of these studies refer to developing countries that have not completed the process of urbanization and where the realities of urban and rural life are clearly poles apart. There is little empirical evidence on similar differences in developed countries.

Spain is particularly suitable for a study of this type as it is characterized by an advanced level of urbanization and development. It possesses a very rich urban structure with several large cities, a large network of medium-sized towns and a rural setting that is still important. Furthermore, differences in earnings have worsened since the onset of the economic crisis and so the breach between high- and low-income households has become wider: the Gini index in Spain increase 2.7 points from 2008 to 2012.

Regional policies oriented to impulse the convergence among territories, urban planning, poverty policies, or programs designed to promote economic growth, productivity or competition should take into account how the consumption patterns and the cost of living change among cities and, in particular, how relevant the effect of the city size might be. Previous research in urban and regional economics has pointed out the existence of substantial differences in costs of living among different sizes of cities, and, also a systematic relationship between the cost of living and the city size has been identified. Most of these studies have been applied for the US, but the number of contributions that analyze this city size effect in Europe is smaller due to data availability and the conclusions less clear. This lack of empirical studies is especially important for the case of Spain, where there is not any quantification of the effect of city size over the cost of living.

The key question asked in this paper is whether the COL is influenced by the agglomerations. The answer is yes and it has been demonstrated through several ways. The first way was the estimation of the COL by municipality size along the in 2012. The results showed that the smallest areas have lower COL consistent with the theoretical and empirical literature revised in previous sections. The difference between the smallest municipalities and the biggest ones is more than 8% in 2012. The second way corroborates the previous results through an alternative approach. In this approach a quantile regression model was used to determine the factors that influence the COL. For this purpose a COL at a microlevel for each household of the HBS has been calculated to regress it over a set of socioeconomic variables and demographic and geographic variables. Among these variables it has been used the cities of more than 100,000 inhabitants to represent the effects of agglomeration over the COL. Through the estimation of a quantile regression it is found that the agglomerations raise the COL but only for the high income quartiles, this result is rational due to the kinds of goods that offers the biggest cities and are only consumed by the rich.

Developing and applying cost of living indicators that allow for spatial comparisons have important policy and welfare implications. Disparities on the average income between large cities and rural or small cities areas (urban premium) could be not as large as they seem if income is adjusted by cost of living differences. Another important implication of not having a proper index of cost of living is the possibility of obtaining misleading results in poverty analysis. A failure to account properly for cost of living differences between urban and rural or small cities areas may lead to regionally inconsistent poverty lines and may result in unwarranted policy interventions. Nominal poverty thresholds that are invariant across space result in an overestimation of the poverty in less urbanized areas compared with urban areas, affecting considerably the eligibility for benefits.

6. References

Atuesta, L., & Paredes, D. (2012). A spatial cost of living for Colombia using a microeconomic approach and cesored data. *Applied Economic Letters, 19*(18), 1799-1805.

Blien, U., Gartner , H., Stüber, H., & Wolf, K. (2009). Regional price levels and the agglomeration wage differential in western Germany. *Ann Reg Sci, 43*, 71-88.

Cebula, R. (1980). Determinants of geographic living cost differentials in the United States: An empirical note. *Land Economics, 56*(4).

Cebula, R. (1989). The analysis of geographic living cost differentials: A brief empirical note. *Land Economics, 65*(1), 64-67.

Cebula, R., & Todd, S. (2004). An empirical note on determinants of geographic living - cost differentials for counties in the State of Florida, 2003. *The Review of Regional Studies, 34*(1), 112-119.

Cooper, R. and McLaren, K. (1992). An empirical oriented demand system with improved regularity properties. *Canadian Journal of Economics, 25*, 652-68.

Deaton, A., & Muellbauer, J. (1980). An Almost Ideal Demand System. *The American Economic Review, 70*(3).

Desai, A. V. (1969). A spatial index of cost of living. *Economic and Political Weekly, 4*(27), 1079-1081.

Haworth, C., & Rasmussen , D. (1973). Determinants of metropolitan cost of living variations. *Southern Economic Journal, 40*(2), 183-192.

Hayes, P. (2005). Estimating UK regional price indices, 1974-96. *Regional Studies, 39*(3), 333-344.

Heien, D., & Wessells, R. (1990). Demand systems estimation with microdata: a censored regression approach. *Journal of Business and Economic Statistics, 8*(3), 365-371.

Hogan, T., & Rex, T. (1984). Intercity differences in cost of living. *Growth and Change, 15*(4), 16-23.

Hoover, E.M. (1937). *Location theory and he shoe and leather industry*. Harvard University Press, Cambridge, MA.

Isard, W. (1956). *Location and Space-economy.* Cambridge, MA: The Technology Press of Massachusets.

Koenker, R., & Basset Jr, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50.

Konus, A. A. (1939). The Problem of the True Index of the Cost of Living. *Econometrica, 7*(1), 10-29.

Kosfeld, R., Eckey, H.-F., & Türk, M. (2008). New economic geography and regional price levels. *Jahrbuch für Regionalwissenschaft, 28*, 43-60.

Kurre, J. (1992). *The cost of liivng in rural Pennsylvania.* Erie, Pennsylvania State University: Center for Rural Pennsylvania.

McMahon, W. (1991). Geographical cost of living differences: an update. *AUREUEA Journal, 19*(3).

McMahon, W., & Melton, C. (1978). Measuring Cost of Living Variation. *Industrial Relations, 17*(3).

Muellbauer, J. (1975). Aggregation, income distribution and consumer demand. *Review of Economic Studies, 42*(4), 525-543.

Nelson, F. (1991). An inter-state cost of living index. *Educational Evaluation and Policy Analysis, 13*(1), 103-111.

Ohlin, B. (1933). *Interregional and Internal Trade.* Cambridge, MA: Harvard University Press.

Serwood, M. (1975). Family budgets and geographic differences in prices levels. *Monthly Labor Review, 98*(4), 8-15.

Shefer, D. (1970). Comparable living costs and urban size: a statistical analysis. *Journal of the American Institute of Planners, 36*(6), 417-21.

Shonkwiler, J. S., & Yen, S. T. (1999). Two step estimation of a censored system equations. *American Journal of Agricultural Economics, 81*(4), 972-982.

Slesnick, D. (2002). Prices and regional variation in welfare. *Journal of Urban Economics, 51*, 446-468.

Timmins, C. (2006). Estimating Spatial Differences in the Brazilian Cost of Living with Household Location Choices. *Journal of Development Economics, 80*, 59-83.

Walden, M. F. (1998). Geographic variation in consumer prices: implications for local price indices. *The Journal of Consumer Affairs, 32*(2), 204-226.

Apendix I. Estimates of the OLS and Quantile Regression over the Expenditure Level provided by the HBS

EXPENDITURE HBS	OLS		QUANTILE REGRESSION									
			10		25		50		75		90	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Cons	8.0740***	167.79	7.2006***	62.31	7.6942***	150.3	8.1733***	166.94	8.5820***	182.85	8.9062***	145.08
Agglomeration	-0.0011	-0.10	0.0017	0.06	-0.0052	-0.32	-0.0064	-0.48	-0.0056	-0.63	-0.0024	-0.22
Income	0.0176	0.32	0.0020	0.02	-0.1008	-1.45	0.0199	0.28	0.0493	1.2	0.1196	1.67
Northwest	-0.0376	-1.21	-0.0280	-0.42	-0.0587	-1.27	-0.0502*	-1.53	-0.0185	-0.61	-0.0253	-1.02
Northeast	-0.0505*	-2.15	-0.0839	-1.26	-0.0400	-1.21	-0.0571	-2.07	-0.0235	-0.92	-0.0572***	-3.46
Central	-0.0139	-0.40	-0.0628	-0.81	-0.0566	-1.14	-0.0237	-0.64	0.0061	0.17	0.0293	0.77
East	-0.0439*	-1.65	0.0105	0.15	-0.0441	-1.21	-0.0772**	-2.76	-0.0308	-1.06	-0.0299	-1.14
South	-0.0179	-0.47	-0.0375	-0.46	-0.0621	-1.35	-0.0347	-0.86	0.0113	0.42	0.0272	0.85
Canary Islands	-0.0100	-0.25	-0.0383	-0.47	-0.0264	-0.59	-0.0150	-0.46	0.0186	0.58	0.0078	0.2
Household Size	0.0126	1.65	0.0420	1.71	0.0138	1.01	0.0123	1.27	0.0019	0.26	-0.0078	-0.76
Number of employed	-0.0126	-1.39	-0.0374*	-1.78	-0.0122	-0.99	-0.0112	-1.45	0.0050	0.93	0.0053	0.55
Age	-0.0002	-0.51	-0.0004	-0.41	0.0003	0.61	-0.0001	-0.3	-0.0002	-0.61	-0.0004	-0.79
Number of dependents	-0.0001	-0.01	-0.0395	-1.35	-0.0004	-0.02	-0.0039	-0.34	0.0046	0.48	0.0073	0.43
First cycle studies	0.0134	0.81	0.0383	0.88	-0.0103	-0.45	0.0133	0.84	0.0057	0.37	-0.0085*	-0.51
Second cycle studies	-0.0064	-0.32	0.0183	0.28	0.0059	0.22	-0.0133	-0.59	-0.0130	-0.73	-0.0459	-1.91
High degree studies	0.0156	0.78	0.0478	1.1	-0.0099	-0.37	0.0198	0.83	0.0079	0.47	-0.0038	-0.18
500-1000 Euros	-0.0038	-0.13	-0.0758	-1.13	0.0041	0.12	0.0133	0.44	-0.0296	-1.23	-0.0113	-0.29
1000-1500 Euros	-0.0100	-0.35	-0.0897*	-2.02	-0.0004	-0.01	0.0052	0.2	-0.0431*	-1.78	0.0014	0.04
1500-2000 Euros	-0.0164	-0.55	-0.0827	-1.38	0.0048	0.13	0.0132	0.47	-0.0630***	-3.05	-0.0187	-0.52
2000-2500 Euros	0.0140	0.44	-0.0386	-0.68	0.0208	0.58	0.0354	1.34	-0.0206	-0.81	0.0145	0.31
2500-3000 Euros	-0.0537	-1.60	-0.1348*	-2.15	-0.0515	-1.39	-0.0375	-1.13	-0.0747**	-2.62	-0.0293	-0.73
More than 3000 Euros	-0.0188	-0.55	-0.0812	-1.41	0.0030	0.07	-0.0045	-0.19	-0.0497*	-1.9	-0.0195	-0.57

Note: *, ** and *** represent the level of significance to 10%, 5% and 1%, respectively.